



下载APP



## 加餐 | 音频技术漫谈之好声音是怎么炼成的？

2021-12-24 冯建元

《搞定音频技术》

课程介绍 >



讲述：正霖

时长 16:02 大小 14.70M



你好，我是编辑冬青。

专栏即将接近尾声，不知道你的学习体验如何呢？可以点击右侧的“评价课程”按钮提供更多的建议给我们哦~我和老师会逐一查看大家的反馈。

作为一堂加餐课，今天我们来学点轻松的，听一场相对简单的分享，看看好声音是怎么炼成的。与建元老师同作为“完美音频”背后默默付出的技术人，我想一定会引起很多共鸣。同时呢，即使脱离音频技术，我想好声音也可以是每个人的标签，不妨看看随着未来音频技术的发展，我们在声音训练上还有哪些可塑的空间吧。



以下是建元老师在课程直播中所做的技术分享，这里以文字的形式分享给你，相关 PPT 以及直播回放可以通过文末的链接获取哦！

## 什么是高音质？

高音质在后疫情时代被大家频繁提及。你可以回想一下，在你的工作和生活中，疫情期间是不是增加了很多实时音频互动场景，那么在后疫情时代，实时音频发生了哪些改变呢？我们一起看一下。

首先，这是一个因果导向的事情，由于疫情催化，我们在实时音频的用法和用量上都有了大幅度的增加，以前你可能不会去做一些在线会议、在线教育等等，现在你都要把这些工具用起来。

在这样的情况下，你对实时音频的质量就会提出更高的要求。比如，日常打电话，可能觉得能听到对方在说什么就 OK 了，但现在你可能会想去看一些直播以及在线音乐场景，这就需要更多的高保真技术支持。这方面在疫情的催化下，其实是有一个飞快迭代的过程的。



实时音频技术在不断的演进，到目前为止，我们的直播、在线 K 歌都可以达到高音质的状态。**那么究竟什么样的音频体验，我们把它定义为高音质呢？**

在实时领域，首先要满足采样率足够高这一条件，要达到 48KHz。简单来说就是，你能听到的大部分声音都能进行采样，这样就不会有频带上的损失。

另一方面，我们现在所使用的一些编解码器能够做到高保真，在 4G、5G 这样普遍存在的情况下，我们可以使用码率较高的音频编解码器，使音频不会因为编解码导致衰减。

然后在平时互动中，不知道你有没有注意到，像一些噪声、回声可能都会对音频产生影响。那么在做处理的时候，其实不可避免的会对音质造成一些损伤，比如你降噪的时候可能把一些音乐也当成噪声被消除了。那么在音乐场景，我们就会做一些低损伤的前处理，尽量使音频保真。

总结来说，**高音质时代就好比开了一个很清晰的摄像头，你脸上的缺点会暴露无遗。**

在高音质的情况下，我们对卡顿、回声这样的质量问题容忍度会下降。比如，以前你打电话，喂喂喂，没有听到，可能就会重播一遍，现在你可能会经常抱怨网络怎么这么差，为什么老是会卡住，为什么老是能听到自己的声音等等，你会对音质更加敏感。

然后在听感上也会追求一些细节。因为打电话的时候采样率只有 8KHz，高频的信息都已经没有了，听不到也不会去追求那些细节。而如果采样率足够了，又是高保真的，说话的时候如果有一些喷麦、杂音等等，那些高频的信息可能就会更加敏感，你可能会去追求能不能体现出这些声音的细节。比如，在一个房间里，能不能体验出空间感。这些都是对音质的新的追求。

## 什么是实时美声？

所以，在这样的情况下，我们就会想能不能让我们的声音更好听呢？当然可以！

比如在录制一首歌曲的时候，就会有调音师帮忙调音，跑调的地方是不是就可以修正一下，喷麦、齿音不好的地方也会被修正。

在实时互动场景中也是一样的。实时互联网发展到今天，在音频或者音视频互动场景中我们已经能够把音质做到无损或者半无损，传递到远端。我们对声音细节会有更加苛刻的要求，就像美颜一样，音频上也会有一些美声。

**实时美声可以让你的声音更好听、更动听。**当然了，在算法设计和应用设计上，我们还要考虑很多问题。比如，在实时互动场景中你去美声，就好比你在实时美颜，如果你的美颜有一些偏差，就会暴露自己的本来面貌。实时美声也是这样的，为了满足实时性以及设备低功耗的要求，在算法设计上我们要考虑设备是不是能把这样的算法跑起来，以及算法带来的额外的延迟会不会导致交流的不顺畅等等。这些在做算法设计之前你就要考虑了。

接下来我们看下，如果你想要做一个实时美声需要做哪些工作呢？我把它称之为**实时美声设计的三驾马车**。这里涉及到的东西会比较多，我在专栏中有做一些知识点的拆解，这里只起到一个先导作用。整体上怎么让声音更好听以及相关细节实现，你可以查看第 12 ~ 14 讲。如果你已经学过这些课程，这里也不妨复习一下。



## 数据驱动

好听是一个非常主观的描述，每个人的定义和取向都不太一样。那么这个时候，我们就要客观定义一下什么样的声音是好声音了。

那在我拿到声音变好听这样的任务时，首先我会去网上找一找相关信息，看看大家觉得什么样的人唱歌是好听的，或者什么样的人的声音是好听的，寻找一些这样的标签。还可以去找一找不同的性别，比如男声、女声分别有什么样的声音是好听的，去找一下相对应的形容词，再给形容词做一个分类。比如，你可能听说，这样的男声非常有磁性，或者这个男人说话十分的稳重，女生可能就说这个人说话很温柔或者比较有活力，这样的形容词就是标签。

在标签的指引下，就会有萝莉音、御姐音这样的分类。加之近些年配音技术的不断成熟，大家对这些名词都很敏感。那像音色方面，可能就会说这个人的音色是比较高亢的、圆润的，其实这些也是标签。

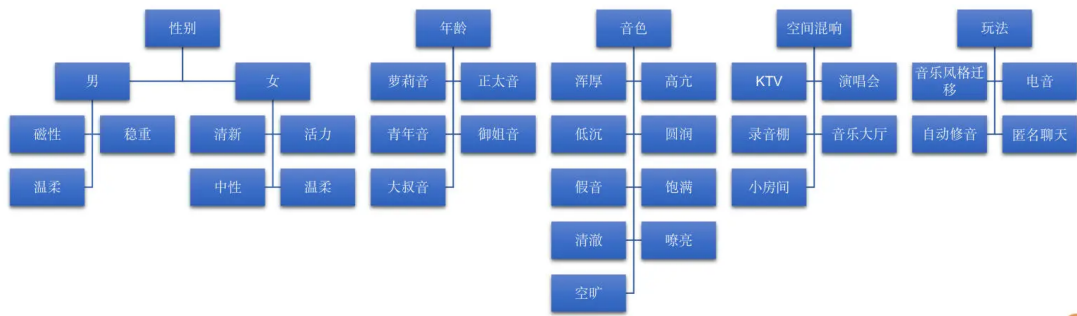
那么除了音色、年龄、性别分类以外，环境也会对声音产生影响。比如某些空间的混响，你在 KTV 唱歌和家里就不一样。再比如演讲会的音乐大厅，这些是专门为大规模的管弦乐或交响乐去做的空间，这种声音很需要空间感的塑造，所以混响也是可以划分为一类的。

而从玩法上来说，声音可以是正常的好听，也可以增加一些好玩的元素。比如用一些电音、自动修音甚至变声，把声音做一些整体的变换，这些都是从玩法上使声音更加好听的一些方向。



## 第一驾马车：数据驱动

在不同的场景收集各种人们所认为的好声音，比如语聊和歌唱，以及不同维度，比如性别、年龄、音色等。建立“好声音”语料库为后续美声算法提供目标。通过数据挖掘将人们的喜好转化为数据与音效设计的目标，用数据驱动音效算法设计。

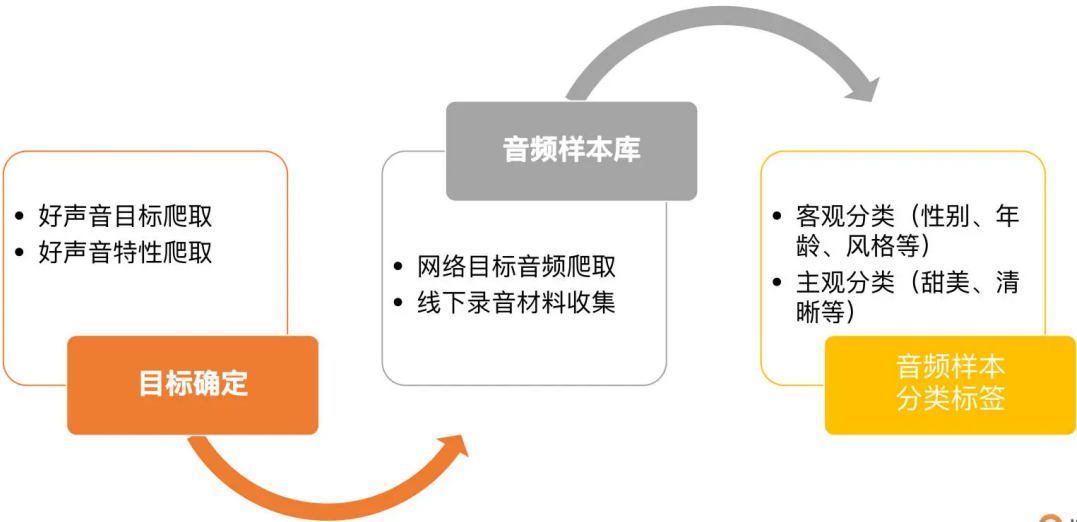


极客时间

有了这样的一堆标签，我们就可以去网上找一些对应的目标，然后把这些声音下载下来，看看能不能收集一些这样的音乐素材。如果你身边恰好有这样的好声音，也可以录制一下，分析一下他们有什么样的特点。从客观上来说，主要是年龄、性别、风格等等，主观上就是上述标签了。**我们可以按照主观、客观这样的大分类来进行数据收集。**



## 第一驾马车：数据驱动



极客时间

## 理论支持

有了这样一些数据样本和数据分类，下一步就是寻找理论支持了。很多好声音，其实你知道它好听，却不知道为什么好听，这就需要很多的理论来帮助我们分析好听在哪里。

## 覆盖的专业领域有很多。

像声学，这个人的发音是不是好，就跟发音腔体有关。比如你的发音是不是足够浑厚，你嘴巴的开度是不是足够大，声音的响度如何，还有共振峰决定了发音的音色，基频决定了发音的音调，这些都是在发声时需要注意的一些方向。声学方面还有混响，刚才提到过，不同房间会有不同的混响。

这块就有很多的理论支持，我们可以提取响度是怎样的，基频是怎样的，混响是怎样的，这样你就可以对声音进行分析了。

另外一块是语言学，主要和韵律、乐理有关。韵律就是指一个人的抑扬顿挫，从指标上来说就是你的音调变化以及你声音响度的变化，有些重读、重音，或者说你这个字拉得特别长，这些像语速、动态调整、语调的变化就是韵律。还是就是音乐上好不好听，就是说你是不是按照正确唱歌的做法去做的，这就跟乐理有关了，比如你是不是在调上，人声和乐器是不是需要配合，不同的音乐风格也会有所不同。

还有像心理学这块，是指我们感知声音。其实声音发成什么样子，在我们心理上的感知又是不一样的，我们可以感知到这个声音是冷色调还是暖色调，是有一些情绪标签的。心理上又可以根据双耳效应来感知你这个声音发的位置，不同的声音它的延迟（比如左右耳）是不一样的，那么在心理上就会感觉出这个位置的方向感也是不一样的。这是我们要在最后做的，让声音从心理上也觉得是好听的。

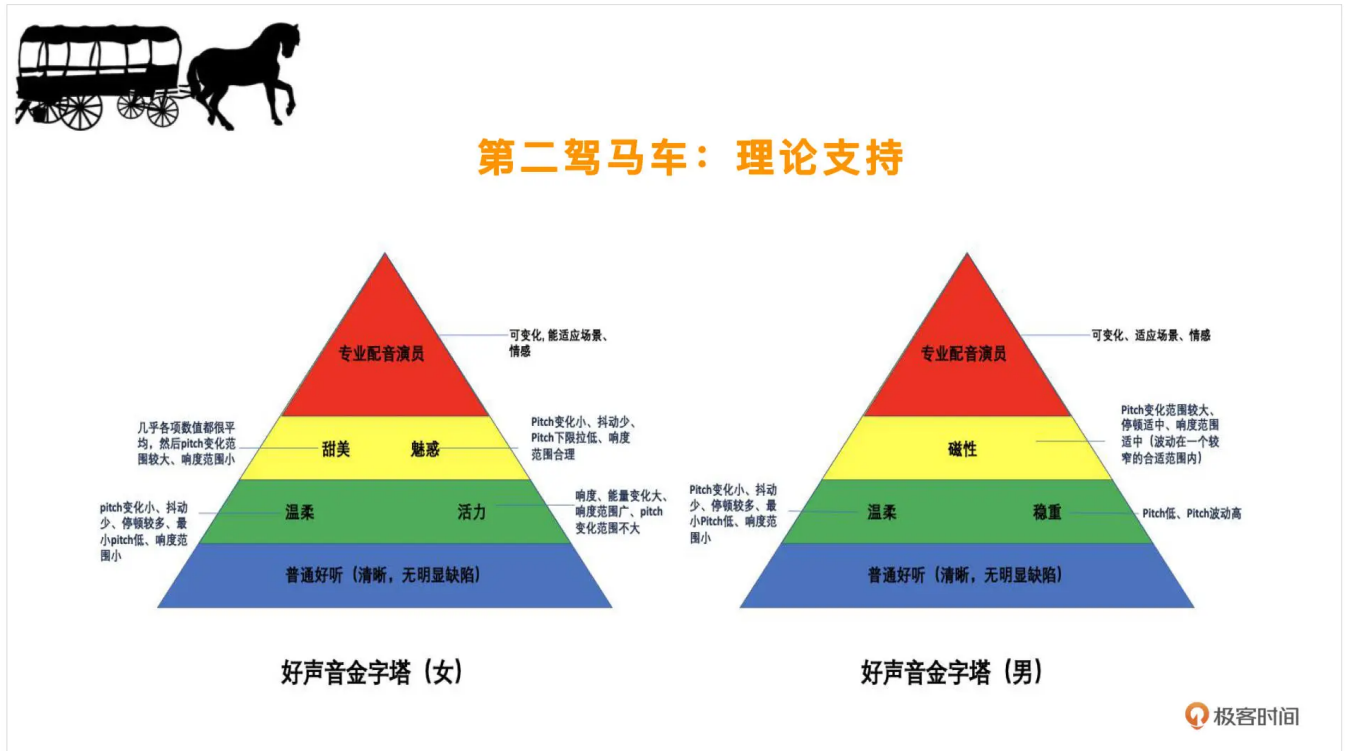
前面是在塑造你发声的器官是不是正确，空间感是不是正确。其中，语言学决定了你的抑扬顿挫是不是正确，乐理决定了你唱歌是不是正确，心理学决定了你听音是不是正确，甚至和你的播放设备也有关系。

这块给了我们很多数学的方式，或者说数学描述的特征，去多维度地分析好声音的一般规律。

举个例子，比如像男性磁性的声音，它其实呈海鸥状，在低频和高频的能量会比较高，中频能量较低，就像一个海鸥展翅的形状，这样的声音往往会表现出比较磁性的特征。再比

如说一些温柔的声音，它的节奏就不会那么快，同时咬文嚼字可能也没有那么清楚，这个时候听上去就会比较温柔。

在有了这些理论支持之后，我们就可以看一下好声音具体是怎么划分的。以下是好声音的金字塔：



首先，我们的好声音一定是分性别的，因为女声的好听和男声的好听其实是不一样的。金字塔的最底端是普通好听，比较明确，这个男女都一样，比如清晰、没有明显的缺陷，这个就是指咬字清楚，没有明显的录音设备导致的缺陷，几乎每个人都能做到。

再往上就需要一些技巧在里面了。像温柔的声音，它的 pitch（音调）变化会比较小，抖动会比较少，停顿会比较多。像一些有活力的声音，男女会各有不同，男的可能稳重点你会觉得好听，但是女生的话，如果你觉得一个女生说话比较磁性，那么不一定是在表扬她，可能她的声音会比较低沉或者沙哑，不一定是好的，所以男女还是要做分类。

再往上就会考验到我們了。假如你是一个配音演员，那你就会需要这层的技巧，普通人可能不一定能发出这样的声音。比如甜美的声音，各项数值会比较平均，但 pitch 变化范围却会比较大，像魅惑还会涉及到一些词语、语言方面的选择，会有更高的要求。

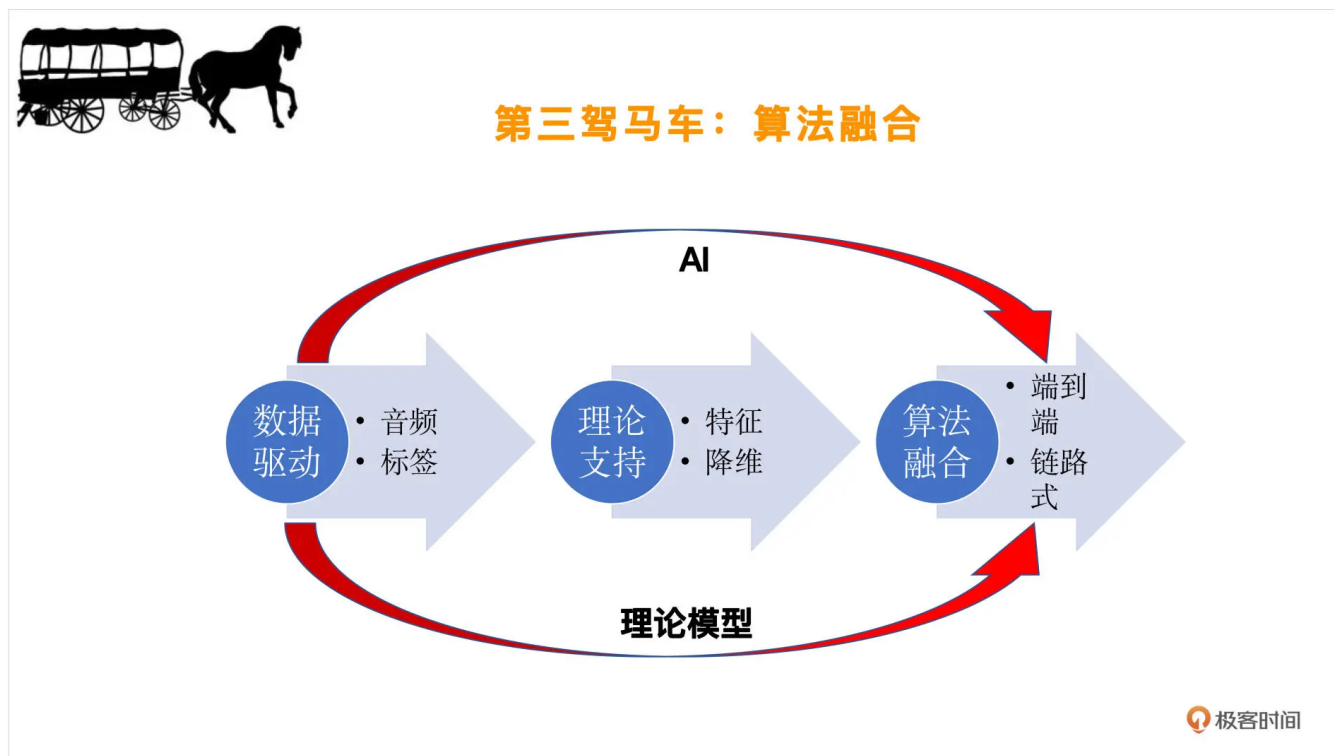
再往上就是专业的配音演员才能达到的层级了。他可以根据不同的场景、不同的情感变化，来自由切换自己的声线，这是最难的，普通人很难实现。

以上就是好声音的金字塔，你可以对照看看自己在哪一层。这里注意一点，这个金字塔只指你修炼的难易程度，也就是自身靠声学训练或者美声训练去做的难易程度，但实际上如果我们用算法去实现，根据不同的场景、情感去做自由切换，则只要有不同的模式可以自由选择就能实现了。

反过来说，算法实现不一定很难，只要有足够的理论支持就可以。

## 算法融合

在我们有了数据驱动——音频和标签，然后根据理论支持明确了好声音的特征之后（哪些特征是重要的筛选一下，我们把它叫做降维），就需要设计算法去调整声音的细节了。



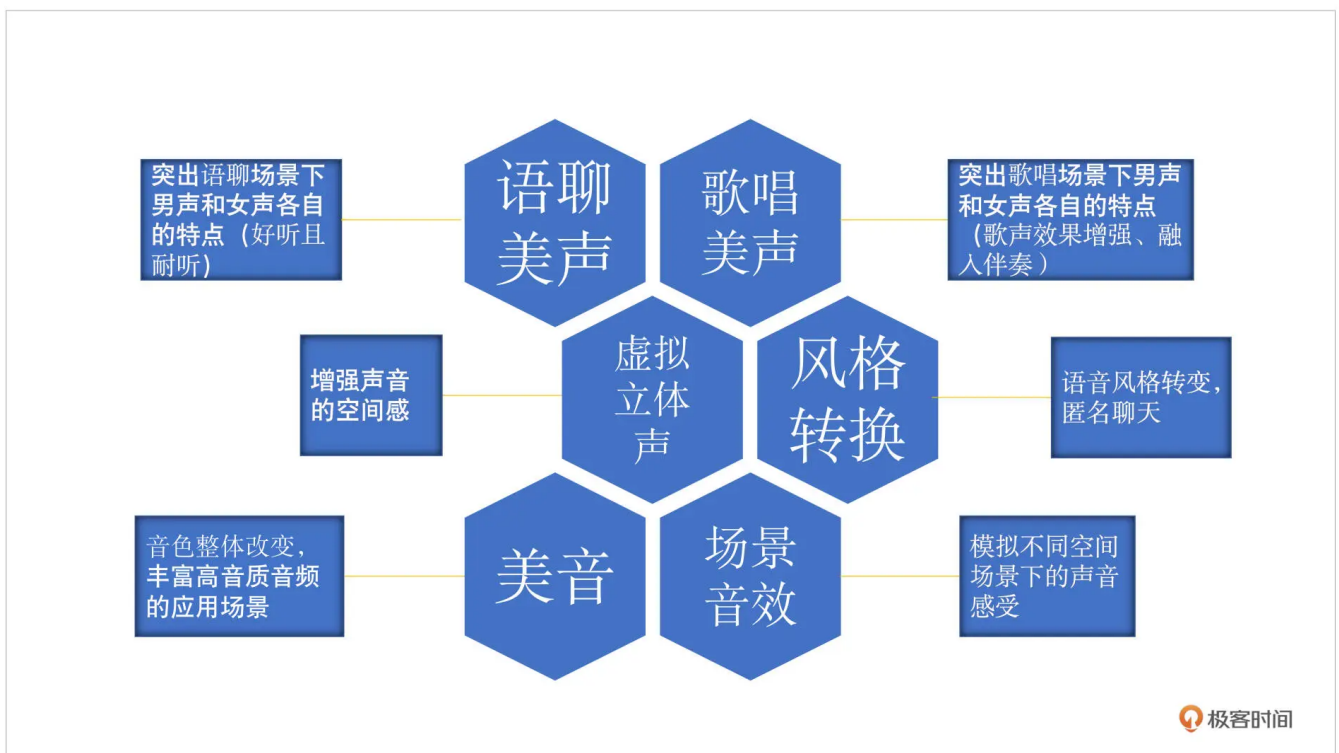
调整的方式主要有这样两种。

第一种，理论模型。比如说我们要去变调，让这个人发音的调性是正确的，就需要做一些修音，这个时候可以采用一些变调算法。然后你觉得混响空间不够贴合伴奏，比如伴奏是在维也纳金色大厅这种比较大一些的音乐会的混响氛围，而这个人唱歌的时候是在混响比

较小的客厅，这时就要改变混响，加入一些混响模型。这种链路式的一个一个模块去改造，根据理论模型就可以实现了，我们也称之为“**链路式的理论模型推导**”。

第二种，端到端的改变。随着 AI 技术的发展，我们可以用一些 AI 的方法，自动提取这个人的风格，而不是提取这样一个一个的链路。比如我们可以做一些整体的变声，整体去改变一个人的音色、音调，以及发音的时长、规律等等。这就是**用 AI 方法做端到端的调整**，这样我们就可以一次性的把这些工作都完成了。

从场景上来说，还会有不同的应用，这里我大概介绍一下会有哪些常用的应用场景，结合以下这张图示一起看一下：



语聊美声主要就是突出男女声不同的声音特点，美声和音效还是有所区别的。语聊美声和歌唱美声主要是在不改变你说话 ID 的情况下，让你的声音更好听。语聊美声会做些小的细微的调整，比如根据你正常的发音，看看有没有受到设备或者自身状态的影响，导致声音不够饱满，什么意思呢？比如你的高频的谐波有很多的缺失，就可以做一些谐波增强或者加一些 EQ 的调整，动态调整一下你的频率范围、频率响度，这里把你的高频提升一些声音就会更加好听。这就是语聊美声。

而歌唱美声则不太一样，它会在歌唱场景下有一个特点，你需要和唱歌、伴奏去融合，刚才提过的混响范围不一样可能会导致唱歌不是那么好听。

音效主要是做一些风格转换，可以用于匿名聊天，这块就可以用一些变声的方法。除此之外还有环境的变换，比如虚拟立体声，我们大部分时候用到的都是单通道的声音，也就是指左右耳发出的声音是一样的，而立体声就是双通道的，通过把单通道变成多通道就可以体现出声音所在的方向感，也就是增强空间感。

那么场景音效，就可以有一些不同的空间场景发出的混响的改变，就像大家普遍会觉得在浴室唱歌特别自信，就是因为加了混响。这和房间的大小、装修材料都是有关系的，都可以通过场景音效模拟实现。

美音主要对应音色，人的音色调整其实是最直观的，比如感冒期间你可能鼻音比较重，高音的部分由于你的鼻腔共鸣被限制就没有了，反而低音部分被加强了，这个时候利用美音把你的 EQ 或者整个平响做一些调整，你的声音就可以从一个感冒的状态变成正常说话的状态。

以上就是这次分享的主要内容，最后还有一些 Demo 试听，直播回放的效果不是最好的，因为它是经过编解码器以及扬声器播放，再回到麦克风去传播，你可以直接打开 PPT 进行播放，再对照老师的讲解，整体体验就非常棒啦！

这里附上直播回放以及 PPT 的获取链接，也期待你能分享更多好玩的音频知识给我们，我们下节课再见！

链接：<https://pan.baidu.com/s/1ITgbKfondObV1dQLtpWhFQ>（密码：und6）

分享给需要的人，Ta 订阅后你可得 **20 元现金奖励**

 生成海报并分享

 赞 0  提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 14 | 音效三剑客：变调、均衡器、混响

精选留言

写留言

由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。